Learning from Data, Assignment Sheet 1

School of Informatics, University of Edinburgh

Instructor: Amos Storkey

Handed out: Mon 16 Oct. Submission Deadline: 23:59 on 2 November 2006.

Please remember that plagiarism is a university offence. Do not show your work to anyone else. The number of marks assigned to each task is given in square brackets. In total, this assignment will contribute 8% to your overall mark for LFD.

The questions ask you to do some MATLAB programming. The routines that you write to accomplish this should be submitted as part of your answer. You are also asked to provide some written answers – these should be written in a plain text file named `answers.txt`. Include your NAME and STUDENT NUMBER at the top of the file, and the LEVEL YOU ARE TAKING THIS COURSE AT (LEVEL 10 or LEVEL 11). Failure to indicate that you are a level 10 student will result in it being marked as level 11. Consider any reduction in mark that results as a penalty for not following instructions!

Do not include markup (e.g. html). Keep your file within 80 characters width for ease of printing. Do not include answers in the code. Make it clear what question you are answering at each point. See `http://anc.ed.ac.uk/~amos/lfd/lfdassignments.html` for submission instructions. Follow the instructions carefully. You should try to make your code clear and comment it. The code will only be looked at if there are any questions regarding the originality of the answers given, or where an unforseen ambiguity arises. Do not include your answers in the code. The marks will be given on the basis of the written work. The code will not contribute to the marks itself.

The marks associated with each question are given. Good insightful answers to questions can serve to increase the baseline marks by at most a further 10 percent up to a maximum of a hundred percent. Verbosity will be penalised. Level 10 students need not attempt the final question, and the marks for each question will be rescaled to total 90 percent. Again 10 percent is available for insightful answers, and the full 100 percent is available to level 10 students without answering the final question. However if a level 10 student does answer the final question, it could contribute to marks for insight.

This assignment is about Data Handling, Principal Component Analysis and Class Conditional Gaussian Modelling.

The data for this assignment is in `assignmentone2006.mat`. This data should be

loaded in to MATLAB using the `load` command, and is to be downloaded by going to the course assignments page: `http://anc.ed.ac.uk~amos/lfd/lfdassignments.html`. Do not attempt to run this file. Download it to your working directory. The `.mat` file contains three datasets in variables `x1`, `x2` and `x3`, where each row is an attribute. In this assignment we are interested in unsupervised learning (the same data will be used for assignment 2 where they will be separated into training and test datasets, and we are interested in predicting the last attribute). This dataset is derived from the `mpg` UCI dataset relating to the miles per gallon of different cars. The last attribute is the miles per gallon a car does, and the other attributes are various features of the car. The three datasets correspond to cars from different regions.

In calculating covariances, please use the form provided by the MATLAB function `cov` throughout, even though this is not strictly the maximum likelihood estimate (it normalises by $N - 1$ not N for number of samples N).

1 [15 Marks]. Write a MATLAB program to perform principal component analysis (PCA) on the data `x1`. You might find it helpful to look at the eigenfaces MATLAB demo, also available from `http://www.anc.ed.ac.uk/~amos/lfd/lfdlectures.html`. What is the largest eigenvalue of the covariance matrix? If you had to choose a number of principal components, what would you choose. And why?

2 [10 Marks] Add to your MATLAB program some code to reduce the `x1` data to the first 2 principal components. With the same matrix and mean that performs the reduction of the training data, reduce the data in the two other datasets (`x2` and `x3`) to 2 dimensions. What is the standard 2 dimensional PCA representation of the first point in each dataset, where the first dimension refers to the largest component, and the second to the next largest? Briefly comment on whether this method for PCA is appropriate, and if not what should have been done and why? Should we have performed the PCA computation separately for each of `x1`, `x2` and `x3`? Justify your answer.

For the purposes of this exercise the PCA representation (but where I have removed the first records of each dataset for obvious reasons!) is available in the file `assignmentonepca2006.mat` from the same website. Please load this data and work with this henceforth, even if you know you have the right answer. Double check the results you have against these results. The data is of the same form, but is now two dimensional.

3 [25 Marks] Learn a mean and covariance for the maximum likelihood Gaussian fit to *each* of the three new datasets containing the PCA representation of the data. Also compute values for the probability that a randomly chosen datapoint comes from each of the datasets. Write down the means, covariances and probabilities you obtain.

We will use these probabilities as priors for a class-conditional Gaussian model. Each different dataset corresponds to a different class. In other words the class corresponds to the origin of the car and we will look at the distributions for each origin of car to try to

predict the origin of any as yet unseen car.

4 [20 Marks] Use these means, covariances and class priors in a class conditional Gaussian model. Write a program to compute the posterior probabilities for any new datapoint. Compute the posterior probabilities for the point $(104, 62, 2327, 229, 200)$ (you will have to reduce this using your PCA computation too). What are you presuming about the generation procedure for this point and for each dataset?

5 [Level 11 only: 20 Marks] By visualising the data and your results, comment on the appropriateness of the class conditional Gaussian model. Comment on how certain you will be on your predictions of class for points coming from the same distribution as the points you have already seen.